Stockage distribué sécurisé pour les sciences humaines et sociales

Joël Marchand

Stockage distribué sécurisé pour les sciences humaines et sociales

Joël Marchand
TGIR Huma-Num - CNRS UMS 3598
54, boulevard
75 006 Paris

Résumé

Huma-Num est une Très Grande Infrastructure de Recherche (TGIR) pilotée par le Ministère de l'enseignement supérieur et de la recherche, et opérée par le CNRS.

Elle rend des services à l'ensemble de la communauté académique en Sciences Humaines et Sociales (SHS) et notamment des services numériques, orientés sur la gestion des données de la recherche, ceci dans le but d'aider les chercheurs à gérer la vie de leurs données.

Huma-Num a conçu et proposé à la communauté un nouveau service de stockage nommé Huma-Num Box et destiné aux gros volumes de données (plusieurs To) dites « froides » ou « tièdes », c'est-à-dire peu accédées et modifiées, mais à forte valeur et donc nécessitant une sécurisation importante.

Il sera exposé la façon dont le projet a été conçu, les objectifs recherchés par rapport à la solution précédente (iRods), la solution retenue (logiciel Active-Circle sur matériel banalisé), les fonctionnalités de la solution, l'architecture réseau mise en œuvre (déploiement sur RENATER au travers de VPN) et l'intégration avec un annuaire LDAP.

Il sera également indiqué les collaborations qui ont été mises en place avec les Maisons des Sciences de l'Homme (MSH), les DSI des universités de rattachement, et RENATER.

Il sera fait un point d'étape sur le déploiement du projet dans 7 points de présence sur le territoire, et indiqué les perspectives d'évolution du projet.

Mots clefs

stockage distribué, stockage sécurisé, archivage, NAS, LDAP, VPN, SHS, Huma-Num Box

1 Contexte et expression des besoins

1.1 Présentation d'Huma-Num

Huma-Num est une Très Grande Infrastructure de Recherche (TGIR) visant à faciliter le tournant numérique de la recherche en Sciences Humaines et Sociales (SHS).

Pour remplir cette mission, la TGIR Huma-Num est bâtie sur une organisation originale consistant à mettre en œuvre un dispositif humain (concertation collective) et technologique (services numériques pérennes) à l'échelle nationale et européenne en s'appuyant sur un important réseau de partenaires et d'opérateurs.

La TGIR Huma-Num favorise ainsi, par l'intermédiaire de consortiums regroupant des acteurs des communautés scientifiques, la coordination de la production raisonnée et collective de corpus de sources (recommandations scientifiques, bonnes pratiques technologiques). Elle développe également un dispositif

technologique unique permettant le traitement, la conservation, l'accès et l'interopérabilité des données de la recherche. Ce dispositif est composé d'une grille de services dédiés, d'une plate-forme d'accès unifié (ISIDORE) et d'une procédure d'archivage à long terme.

La TGIR Huma-Num propose en outre des guides de bonnes pratiques technologiques généralistes à destination des chercheurs. Elle peut mener ponctuellement des actions d'expertise et de formation. Elle porte la participation de la France dans l'infrastructure européenne DARIAH en coordonnant les contributions nationales.

La TGIR Huma-Num est portée par l'Unité Mixte de Services 3598 associant le CNRS, l'Université d'Aix-Marseille et le Campus Condorcet.

1.2 Stockages existants

La grille de services numériques opérée par Huma-Num est basée sur une infrastructure située au Centre de Calcul de l'IN2P3 (CC-IN2P3) à Villeurbanne près de Lyon. Elle propose plusieurs services de stockage sécurisé de données reposant sur deux systèmes informatiques sur disques :

- le système utilisant le logiciel libre iRods, opéré par le CC-IN2P3, dont Huma-Num est l'un des clients pour environ 130 To ;
- un système NAS de marque NetApp, accueillant
  - les données pour l'hébergement de sites Web et les applications SIG,
  - les disques virtuels et les volumes de données pour les serveurs virtuels,
  - un gestionnaire de fichiers à interface Web (logiciel commercial FileRun),
  - un dispositif interopérable de partage de données associées à des métadonnées basé sur les technologies du web sémantique (système NAKALA conçu et développé par Huma-Num)

  pour un volume de 50 To de données à ce jour.

Ces deux systèmes ont chacun des limitations :

- pour iRods : interface d'usage en ligne de commande (CLI) très peu adaptée aux utilisateurs de notre communauté SHS ; impossibilité de mettre en œuvre une application classique sur ces données sans réécriture du code accédant aux données ; difficulté à gérer plusieurs versions d'un même fichier ;
- pour les services proposés sur le système NetApp : difficulté à manipuler des grands volumes de données (uniquement possible par WebDAV dans le cas de FileRun) ; pas d'ouverture possible du système sur Internet autrement que par une interface Web ou SFTP (les protocoles de partages de fichiers comme NFS et SMB étant limités à des usages sur des réseaux locaux (LAN)).

Dans les deux cas, nous ne disposons pas de fonctions NAS traditionnelles, accessibles depuis le poste de travail des chercheurs répartis sur le territoire. De plus, la sécurisation de ces systèmes, *via* le logiciel de sauvegarde TSM et les deux grandes robotiques du CC-IN2P3, ne permet pas d'avoir une vision unifiée des données (sur disques et sur bandes).

1.3 Présentation du projet

Une partie importante de la communauté nationale de recherche en SHS se situe au sein des Maisons des Sciences de l'Homme (MSH). Ces structures ont le label Infrastructures de Recherche. Au nombre d'une vingtaine, elles sont réparties dans les régions et sont abritées au sein des établissements universitaires. Elles sont donc reliées au réseau national de la recherche RENATER.

Fin 2015, Huma-Num a décidé de compléter son offre de service par un dispositif de stockage distribué au sein des MSH et sécurisé sur ses points de présence à Villeurbanne et Paris. Ce service visait à faciliter pour les chercheurs le stockage, la sécurisation et la gestion de leurs jeux de données.

1.4 Caractéristiques des données

Le nouveau dispositif vise à stocker les données réputées « tièdes » voire « froides » au sens où peu voire très peu d'accès en écriture comme en lecture seront effectués durant toute la vie de ces données. En revanche, ces données ont vocation à être conservées de manière fiable durant plusieurs années (5 à 10 ans), car elles constituent la matière première du travail des chercheurs et ont souvent une valeur de type patrimonial. Le service de stockage peut être vu comme le pendant numérique d'une armoire sécurisée où l'on stocke des documents importants, à la différence d'un bureau où se trouvent les documents courants et de toutes natures.

Ces données sont notamment issues de campagnes de numérisation de fonds anciens, de photos, d'enregistrements audio, de cartes, de vidéos, voire de modèles 3D. Elles sont uniquement sous forme de fichiers, souvent de grande taille (mais pas exclusivement), parfois accompagnés de fichiers de métadonnées. Leur volume peut atteindre plusieurs To voire dizaines de To par jeu de données.

Ces jeux de données sont gérés par des équipes de recherche, parfois situées sur un même site mais parfois réparties sur plusieurs. Ces données peuvent être mises à disposition en lecture à un cercle plus important de chercheurs, puis souvent en libre accès sur Internet. Au sens informatique du terme, il est donc recherché la possibilité de constituer des partages pour chaque jeu de données, disposant de règles d'accès comme celles utilisées sur des services de fichiers classiques.

Enfin, ces données ne présentent pas de sensibilité particulière en termes de confidentialité. En revanche, leur intégrité doit être maintenue sans faille au cours du temps, durant toute leur vie.

1.5 Caractéristiques fonctionnelles du dispositif recherché

Le dispositif recherché devait amener un service de type NAS (NFS, CIFS, avec authentification sur un annuaire Active-Directory/LDAP global) au sein des réseaux locaux des MSH.

Ceci afin de donner aux utilisateurs :

— de la souplesse : par opposition aux dispositifs de filtrage sur Internet dans le cas d'accès à une ressource distante ;

— de la liberté : dans le choix des outils de manipulation et des applications utilisant cette ressource de stockage ;

— des temps de réponses courts.

Pour chaque jeu de données, il était demandé de pouvoir définir facilement et de faire évoluer dans le temps la sécurité appliquée sur ces données :

— personnes pouvant accéder en lecture seule ou en lecture-écriture à ces données ;

— sites depuis lesquels ces données sont accessibles (sans forcément qu'une copie y soit présente) ;

— nombre d'instances (entre le site producteur et les deux sites centraux d'Huma-Num), pouvant varier typiquement de 1 à 3 ;

— supports utilisés (si possible en sachant mixer disques et bandes magnétiques) ;

— gestion des versions dans le temps ;

— mécanisme paramétrable de rétention permettant de conserver des données supprimées par les utilisateurs (mais restant accessibles aux administrateurs du dispositif).

De plus, nous souhaitions disposer des fonctionnalités suivantes :

— possibilité, pour un jeu de données ayant plusieurs instances sur plusieurs sites, que la modification puisse se faire depuis tous les sites (synchronisation multi-directionnelle) ;

— possibilité depuis une MSH d'injecter des données dans le dispositif, sans conservation locale d'une instance sur son site ;

— fonctions de stockage sous forme purement logicielle, en utilisant du matériel banalisé ;

— vision unifiée sous forme d'un catalogue global, afin de permettre une administration simplifiée de tous les jeux de données, de toutes les versions, et ceci au cours du temps ;

— fonction d'archivage, c'est-à-dire capacité à figer à un moment donné un jeu de données et de le protéger contre toute possibilité humaine ou accidentelle d'effacement.

2 Choix, configuration, intégration, déploiement situation actuelle

2.1 Choix de la solution

À partir de l'expression de besoins précédente, un appel d'offres a été lancé à l'automne 2015. Trois réponses ont été reçues :

— solution Microsoft avec du matériel HP basée notamment sur DFS et un logiciel d'indexation et de fouille de données : cette solution ne présentait pas les fonctionnalités de sécurisation recherchées ;

— solution NetApp avec des baies NAS et les fonctionnalités SnapVault et SnapMirror : la réputation de cette solution était excellente, l'offre était la mieux disante en termes de volumétrie nette utile, mais l'ensemble des fonctionnalités recherchées n'était pas présent ;

— solution logicielle Active-Circle sur du matériel Dell : cette solution présentait les avantages d'apporter toutes les fonctionnalités recherchées, d'avoir une architecture où les parties matérielle et logicielle sont indépendantes, et d'être proche quantitativement de la solution NetApp.

Cette dernière solution a été retenue.

2.2 Présentation et configuration de la solution

Le logiciel Active-Circle, développé depuis une quinzaine d'années par la société française du même nom et rachetée récemment par une autre société française (Oodrive) est une solution de stockage et d'archivage de fichiers. Elle permet de virtualiser un ensemble de ressources physiques (disques et bandes) rattachées à un ensemble de serveurs, pour constituer un unique ensemble de stockage, présenté sous forme d'interfaces NAS (NFS, CIFS, FTP). Elle permet de définir des classes de services sur les partages (jeux de données), qui concernent le nombre et la localisation de leurs instances (copies d'un jeu de données), la politique d'historisation des fichiers, la politique d'archivage sur des bandes.

La configuration acquise suite à l'appel d'offres a été rapidement étendue pour atteindre à ce jour la configuration suivante :

— 7 serveurs Dell R730 avec 64 Go de mémoire, 2 disques SAS en RAID 1 pour le système, 8 disques SATA de 1 To en RAID 10 pour les catalogues ;

— 7 baies MD1400 en attachement SAS comportant 12 disques SATA de 6 To pour le stockage des données, ce qui représente 55 To nets utiles ;

— les licences Active-Circle pour 7 nœuds, 300 To de stockage utile sur les baies de disques, les extensions API REST, checksum et AME.

Ceci est illustré par la figure suivante.

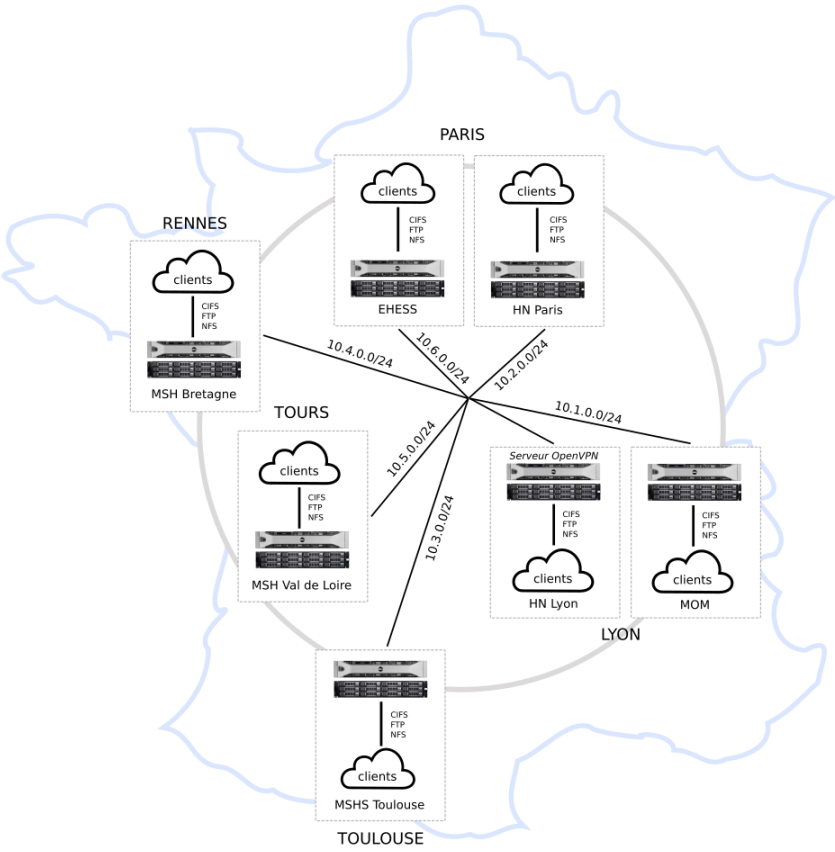

Figure 1 - Architecture actuelle

2.3 Intégration

À cela, il a fallu adjoindre deux autres composants logiciels.

Tout d'abord, chaque nœud Active-Circle a besoin de communiquer de manière bi-directionnelle avec de nombreux ports de tous les autres nœuds. Aussi, dans le cadre d'un déploiement WAN sur divers sites en France, il était indispensable de disposer d'un réseau VPN d'interconnexion entre ces nœuds. Une étude sur le logiciel OpenVPN a été menée, afin d'optimiser au mieux sa configuration, pour tirer parti des liens 1 Gb/s reliant les nœuds. Les détails concernant cette configuration sont à disposition auprès de l'auteur.

Ensuite, comme pour tout serveur de fichiers proposant des services NAS, il fallait coupler le logiciel Active-Circle à un annuaire permettant une authentification pour les protocoles NFS (uid/gid), CIFS et FTP (login + mot de passe). Ne disposant pas d'un tel annuaire au sein d'Huma-Num, il a été décidé de mettre en œuvre une architecture OpenLDAP redondée (1 maître et 2 réplicats) et sécurisée (uniquement sur SSL et avec authentification systématique pour toute requête), avec le logiciel libre FusionDirectory comme frontal de gestion Web. Grâce à une prestation auprès de la société Opensides, contributeur principal à FusionDirectory, cette architecture a été mise en place, sécurisée, documentée, et paramétrée dans un but de délégation. Il apparaissait nécessaire de donner le maximum d'autonomie aux sites qui allaient accueillir un nœud de stockage pour gérer les utilisateurs et les groupes associés à des partages. FusionDirectory a permis de réaliser cela, tout en amenant une interface Web d'appropriation aisée par des personnes ne connaissant pas du tout LDAP. Les détails concernant cette configuration sont à disposition auprès de l'auteur.

Aussi l'intégration de la solution a consisté principalement en :

— l'optimisation de l'usage d'OpenVPN ;
— la mise en œuvre et la prise en mains de FusionDirectory ;
— la recherche avec l'éditeur d'ActiveCircle du bon paramétrage du logiciel, pour la prise en compte de la configuration réseau où les nœuds doivent communiquer entre eux, non pas par toutes leurs interfaces réseau, mais uniquement par leur interface VPN ;
— l'attente vis à vis de l'éditeur d'Active-Circle d'une évolution du logiciel pour sa capacité à réaliser une authentification pour un accès CIFS à partir d'un serveur OpenLDAP utilisé comme PDC (et non pas un serveur Active-Directory).

2.4 Déploiement

L'opération de déploiement a été nettement plus longue que prévue.

Nous avons commencé par identifier des Maisons des Sciences de l'Homme (MSH), volontaires pour être sites pilotes et pouvant accueillir une telle solution. Un document détaillé de présentation de la solution, de ses capacités, de ses limitations, et des prérequis techniques d'hébergement a été rédigé. Il a permis d'amorcer un échange avec la direction des MSH et les personnes pouvant devenir nos correspondants fonctionnels et techniques.

Dans la plupart des cas, ces correspondants techniques ont été des ingénieurs des Directions des Systèmes d'Information (DSI) des universités où se trouvent les MSH. Nous avons expliqué et motivé la logique non « standard » de leur demander d'accueillir dans leur datacentre un serveur qui ne leur appartiendrait pas et dont ils n'auraient pas la gestion.

Les correspondants fonctionnels ont diverses origines professionnelles et une expérience variée de l'outil informatique, et notamment des serveurs de fichiers. De nombreux échanges ont eu lieu afin de les aider à s'approprier l'interface de gestion des comptes (FusionDirectory), son articulation avec le logiciel Active-Circle, la co-gestion des partages entre eux et Huma-Num. Tout ceci dans un contexte d'adaptation au fil du temps de la configuration du dispositif, qui est passé au cours de l'année 2016 du statut expérimental à la production.

Huma-Num remercie vivement toutes les personnes avec qui nous avons interagi durant cette année 2016, car elles ont été patientes, à l'écoute et très constructives dans leurs questions, remarques et demandes. C'est en grande partie grâce à elles que le projet a pu aboutir.

Une fois l'ensemble des prérequis réunis (capacités d'hébergement, appropriation et test de la solution, engagement à utiliser la solution à une hauteur de plusieurs dizaines de To), nous avons procédé à l'envoi des serveurs et des baies dans les MSH, *via* l'unité de service du CNRS spécialisée dans le domaine (Ulisse).

Afin d'illustrer au mieux le rôle du service proposé aux MSH, nous avons désigné chaque nœud du dispositif par le terme Huma-Num Box.

2.5 Situation actuelle

À ce jour, nous disposons donc d'une solution en production depuis environ un an, comportant 7 serveurs répartis sur RENATER (Lyon (2), Paris (2), Rennes, Toulouse, Tours). Outre les deux nœuds de consolidation dans les deux points de présence d'Huma-Num, cinq MSH ou établissements disposent d'un nœud dans leur réseau local.

L'accès pour les utilisateurs raccordés aux réseaux locaux où se trouvent les MSH disposent donc des protocoles CIFS, FTP et NFS, pour accéder aux partages. Ceci comme avec n'importe quel serveur de fichiers.

De plus, tous les utilisateurs référencés dans l'annuaire, et notamment ceux situés sur des réseaux locaux autres que ceux où se trouvent les nœuds, disposent d'un accès SFTP (au-dessus de SSH), *via* une VM hébergée à Huma-Num, qui ouvre son port SSH sur Internet, authentifie ses utilisateurs sur l'annuaire LDAP, et accède aux partages Active-Circle par NFS.

Enfin, les premiers essais d'usage de CIFS *via* la connexion du poste de travail au VPN d'interconnexion des nœuds sont fort encourageants. Ainsi l'accès à son disque réseau devient possible en situation de mobilité.

L'inconnue technique principale du projet était le bon fonctionnement et la stabilité de nœuds Active-Circle sur une architecture WAN. Ceci n'avait jamais été testé. Il s'avère clairement qu'aucun souci n'est identifié à ce jour. *A contrario*, nous avons mesuré des vitesses de duplication site à site saturant les liens 1 Gb/s entre deux nœuds au travers de RENATER.

Plus de 50 partages sont définis, dont plus de 30 sont alimentés (entre quelques dizaines de Go et 40 To).

Au total, ils représentent 210 To de données utilisateurs, stockées pour certains partages avec une seule instance dans le dispositif (car le fournisseur des données dispose d'un stockage primaire par ailleurs), et pour les autres avec deux instances (typiquement soit une instance dans une MSH et une instance dans un des 2 sites Huma-Num, soit dans les 2 sites d'Huma-Num). Vu des baies de stockage et de la licence Active-Circle, 223 To sont occupés.

Fonctionnellement, nous constatons que la communication régulière autour du projet et l'accompagnement personnalisé auprès de chaque correspondant fonctionnel ont permis une bonne appropriation du fonctionnement et des usages possibles de la solution. Cependant, il s'avère qu'il faut plus de temps que prévu initialement pour que les chercheurs sortent leurs jeux de données de leur tiroir (DVD, disques amovibles) et les basculent sur une telle solution. Nous espérons que la stabilité et les bons échos autour du dispositif vont contribuer à dynamiser ces usages.

Il est à noter que nous avons accueilli et allons continuer à accueillir de nombreuses demandes émanant de laboratoires ou de projets de recherche possédant depuis longtemps un serveur de fichiers faisant office de NAS en leur sein, mais ne disposant pas d'un lieu externe à leur campus pour dupliquer leurs données afin de les sécuriser.

Enfin, nous avions envisagé des performances du dispositif, en termes de vitesse d'entrée/sortie et de temps d'accès aux fichiers, assez modestes, vu la complexité des traitements opérés par Active-Circle pour traiter un fichier (découpé en paquets élémentaires de 32 Mo) et maintenir à jour les nombreuses informations du catalogue du partage où est situé le fichier. Divers tests de transferts (typiquement par rsync) sur des ensembles de fichiers de taille importante (plusieurs Go) ont permis de mettre en évidence que des échanges y compris en écriture à plus de 800 Mbs sont possibles.

Aussi nous avons commencé à expérimenter un usage qui était très hypothétique au départ, consistant à utiliser le dispositif comme stockage directement utilisé par une application Web pour publier des médias en ligne. Pour un premier site comportant des contenus audio, nous avons ainsi basculé les 7 To de médias de notre NAS NetApp sur un partage Active-Circle, sans que les consultations de ces medias soient ralenties de manière perceptible par l'internaute.

3 Perspectives

3.1 Extensions

Les bons résultats de cette première phase du projet incitent à la fois Huma-Num et la communauté SHS à étendre l'usage du dispositif actuel.

Ainsi deux autres nœuds seront prochainement mis en œuvre à Nantes et Nanterre (où 65 To de données sont d'ores et déjà en attente). Plusieurs laboratoires nous ont aussi demandé de dupliquer leurs données, pour une volumétrie d'environ 100 To.

Par ailleurs, nous sommes en phase de décision pour adjoindre aux nœuds actuels constitués de baies de disques, un nœud pilotant une robotique de bandes LTO-7. Les motivations sont les suivantes :

— une double sécurité : les risques liés aux stockages sur disque ou sur bande sont très différents ;

— une optimisation des coûts de stockage : une instance d'un jeu de données sur bandes va vraisemblablement coûter environ moitié prix par rapport à une instance sur disque ;

— un premier niveau d'archivage *bit à bit* : Active-Circle permet de déclencher de manière programmée des copies d'un partage sur un jeu de bandes dédiées, qu'il est ensuite possible de verrouiller logiquement et physiquement. Ceci permet d'accroître encore la sécurisation des données vis à vis des rançongiciels, des erreurs humaines de manipulation ou des bogues ou corruptions majeures causées par l'ensemble des logiciels utilisés (firmwares des disques et des contrôleurs RAID, logiciel et catalogues Active-Circle). En effet l'écriture, sous les formats standards TAR ou LTFS de ces archives sur bandes rend leur relecture possible, indépendamment d'Active-Circle lui-même.

3.2 Passage de OpenVPN au service L3VPN de RENATER

L'usage actuel de OpenVPN entre les nœuds comporte plusieurs inconvénients :

— Une organisation en arbre : la communication entre deux nœuds passe forcément par le serveur OpenVPN à Lyon, même si les deux nœuds sont à Tours et Rennes ;

— Un point de faiblesse : si le serveur OpenVPN n'est pas disponible, aucun nœud ne peut communiquer avec aucun autre ;

— Un coût lié au chiffrement : dans la configuration actuelle, un mécanisme de chiffrement des données est activé. Malgré l'optimisation faite, lors d'un échange voisin de 1 Gb/s, cela consomme un cœur de CPU sur chacune des deux extrémités. Cela pourrait finir par charger considérablement le serveur OpenVPN lors d'échanges simultanés entre plusieurs nœuds.

Aussi nous avons entamé depuis plusieurs mois les démarches auprès de RENATER et des sites concernés, pour passer sur le service L3VPN de RENATER. Celui-ci apportera une réponse positive à tous les points précédents. Nous sommes en train de procéder à la mise en place technique de cette solution sur les sites.

3.3 Inconvénients et limites

Nous avons déjà identifié un certain nombre d'inconvénients et de limites au dispostif en place.

Les performances des services NAS sont plus modestes qu'avec un serveur ordinaire. Comme expliqué précédemment, cela n'est pas forcément un handicap majeur, car la cible reste des données peu accédées, et avec quasiment aucun accès concurrent.

Il n'y a aucune limite technique en termes de nombre de nœuds, de volumétrie ou de nombre de fichiers, qui pourrait nous impacter : des installations d'Active-Circle à plus de 50 nœuds, plusieurs Po de données et plusieurs centaines de millions de fichiers existent par ailleurs.

À ce jour, le service CIFS proposé implémente uniquement le protcole SMB version 1. Un contournement envisagé à court terme est de mettre un service CIFS porté par un serveur Samba version 3, qui servirait des clients CIFS en SMB version 2, tout en étant lui-même client NFS d'Active-Circle.

L'éditeur d'Active-Circle ne s'engage pas à ce jour sur plus de 250 partages sur l'ensemble des nœuds. Cela obligera à maitriser le nombre de partages par site, au fur et à mesure de la croissance du nombre de sites. Une solution est de jouer sur les droits au niveau du filesystem (POSIX et NTFS), en plus des droits d'accès et des caractéristiques des partages au niveau d'Active-Circle. Il est ainsi possible de subdiviser un seul partage en sous-dossiers ayant des droits d'accès limités à des groupes LDAP donnés.

Une autre limite intrinsèque d'Active-Circle est le nombre maximum de fichiers ou de sous-dossiers dans un dossier donné. Ceci vient des mécanismes internes de traitement du logiciel. Aujourd'hui cette limite est fixée à 32000, ce qui est parfois atteint sur des dossiers contenant plein d'objets identiques que le producteur des données n'a pas jugé utile ou nécessaire de répartir en plusieurs sous-dossiers. Un échange au cas par cas est à envisager avec lui s'il peut réorganiser son arborescence.

Active-Circle est un logiciel commercial, porté par un éditeur de taille moyenne, ayant une base installée assez réduite et restreinte en grande partie à la France. Même si de grosses institutions de l'enseignement supérieur et de la recherche comme le Synchrotron Soleil, l'Observatoire de Paris, l'IGN ou l'ICM l'utilisent, cela représente une forme de fragilité. Néanmoins la capacité de pouvoir ressortir les données via des protocoles standards et à recycler le matériel banalisé (serveurs X86 et baies SAS passives) constituent des éléments rassurants, si demain pour une raison ou une autre, nous étions motivés à changer de solution de stockage. À notre connaissance cependant, il n'existe pas à ce jour de solution ayant toutes les caractéristiques d'Active-Circle, et ayant fait preuve d'une bonne stabilité et fonctionnant sur des réseaux longue distance.

Le dispositif actuel est internalisé au sein de l'unité de services Huma-Num, qui en devient responsable. Par contre, l'ensemble des données reste dans une organisation strictement interne à la communauté académique, et même du secteur des sciences humaines et sociales, qui pourra dans l'avenir en faire évoluer s'il le souhaite l'organisation.

Enfin nous sommes bien conscients que ce dispositif reste un dispositif purement technique d'infrastructure. Comme toute autre solution et indépendamment des technologies de stockage utilisées, ce dispositif ne sera un succès que par une communication, un accompagnement et un relai continus vers les chercheurs sur chaque site. En effet les chercheurs en sciences humaines et sociales n'ont pas pour la très grande majorité d'entre eux de culture et de pratique en termes d'usage, d'organisation et de gestion d'un grand volume de données. Rendre l'usage d'un tel dispositif le plus simple possible (raccourcis sur le bureau), former à l'organisation des arborescences, insister sur le besoin de documenter ses données avec des métadonnées, organiser dans le temps la conservation et le référencement de ces données, mettre en œuvre les bonnes applications au-dessus de ces données restent des objectifs fondamentaux.

3.4 Développement des usages

Le renforcement du niveau de sécurité du stockage proposé par l'ajout de bandes magnétiques nous permet d'envisager de proposer à nos utilisateurs un service de pré-archivage, en amont de celui d'archivage pérenne proposé par l'opérateur national de notre communauté, à savoir le CINES. Celui-ci propose une plateforme complète d'archivage numérique conforme au modèle OAIS, bien plus riche fonctionnellement qu'une simple conservation d'un ensemble de fichiers. Nous renvoyons le lecteur au site du CINES pour la description de ce service. Huma-Num pourrait donc utiliser l'architecture et les fonctionnalités d'Active-Circle pour stocker un certain temps les jeux de données et de métadonnées, le temps que les scientifiques producteurs de ces données, Huma-Num et le CINES, se soient accordés pour le versement au CINES.

Comme évoqué précédemment, nous avons dès à présent de nombreuses confirmations de développement des usages (remplissage et nombre de partages), à la fois par les MSH disposant d'un nœud Active-Circle, et par des entités souhaitant dupliquer leurs données, voire les mettre à disposition d'autres équipes scientifiques.

De plus, nous allons vraisemblablement basculer les médias de deux autres sites Web importants hébergés au sein d'Huma-Num sur ce dispositif. Cela permettra de valider la capacité d'Active-Circle à servir au travers d'applicatifs Web des volumes importants de données, peu voire très peu consultées, tout en assurant leur sécurisation.

Par ailleurs, les autres fonctionnalités techniques proposées par Active-Circle comme la gestion des checksums des fichiers (conservation d'une signature de référence, et calcul à la demande pour comparaison

entre la signature courante et celle de référence), une API REST permettant d'écrire des requêtes vers Active-Circle depuis une application externe, et un explorateur de fichiers (AME pour Active-Circle Media Explorer, basé sur le logiciel libre ResourceSpace) seront aussi explorées.

En conclusion, un des objectifs initiaux du projet pour Huma-Num qui consistait en la mise en place d'un réseau de stockage sécurisé de données de recherche en sciences humaines et sociales prend forme. En effet, plus le dispositif va se développer, plus il va permettre de donner accès à travers RENATER à des jeux de données, à un chercheur autorisé par les producteurs des données, et ceci via des protocoles d'accès simples d'emploi, même si le chercheur ne se trouve pas sur un site où sont stockées ces données.